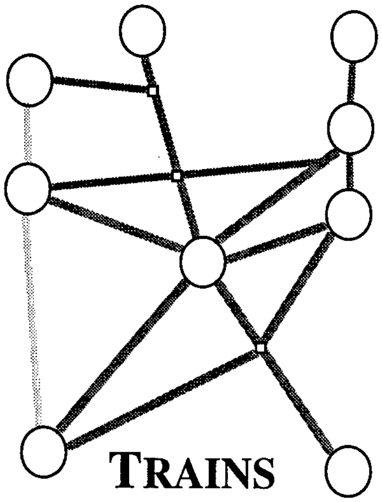

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

TRAINS-96 System Evaluation

Amanda J. Stent and James F. Allen

TRAINS Technical Note 97-1
March 1997

19971007 128

DTIC QUALITY INSPECTED 4

UNIVERSITY OF
ROCHESTER
COMPUTER SCIENCE

TRAINS-96 System Evaluation

Amanda J. Stent and James F. Allen

March 19, 1997

Abstract

In this report we describe an experiment designed to:

- evaluate the performance of the TRAINS-96 system as a whole
- examine the utility of a new robust post-parser module, recently added to the TRAINS system
- explore the benefit to the user of receiving system feedback on speech input

The evaluation uses the same task-based methodology as was used for the TRAINS-95 evaluation [7], in which the user and computer cooperatively solve a given problem. Success is measured in terms of task performance measures such as time to completion of a task, and the quality of the final plan produced.[1]

[1] This work was supported in part by the Office of Naval Research under research grant no. N00014-95-1-1088, and by the National Science Foundation under grant no. IRI-9623665.

REPORT DOCUMENTATION PAGE

Form Approved

OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE
March 1997

3. REPORT TYPE AND DATES COVERED
technical report

4. TITLE AND SUBTITLE

TRAINS-96 System Evaluation

5. FUNDING NUMBERS

ONR N00014-95-1-1088

6. AUTHOR(S)

A.J. Stent and J.F. Allen

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESSES

Computer Science Dept.
734 Computer Studies Bldg.
University of Rochester
Rochester NY 14627-0226

8. PERFORMING ORGANIZATION

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESSES(ES)

Office of Naval Research
Information Systems
Arlington VA 22217

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

TRAINS TN 97-1

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Distribution of this document is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

(see title page)

14. SUBJECT TERMS

TRAINS; spoken dialogue systems; task-based evaluation; natural language processing

15. NUMBER OF PAGES

40 pages

16. PRICE CODE

free to sponsors; else $2.00

17. SECURITY CLASSIFICATION OF REPORT
unclassified

18. SECURITY CLASSIFICATION OF THIS PAGE
unclassified

19. SECURITY CLASSIFICATION OF ABSTRACT
unclassified

20. LIMITATION OF ABSTRACT

UL

1 Introduction

TRAINS-96 is an extension of the TRAINS-95 system, part of "a long-term effort to develop an intelligent planning assistant that is conversationally proficient in natural language" [7]. The domain is a train route planner, where a human manager and the system cooperate to develop and execute plans [3, 2]. The user is able to interact with the system by clicking on objects with a mouse, selecting items from menus, or speaking or typing to the system in English. The system interacts with the user using spoken and displayed English and through graphical displays.

The TRAINS system is designed to help researchers implement and test computational theories of natural language, dialogue and planning. The TRAINS-96 system builds on the TRAINS-95 system, adding realistic distances and times, and allowing users to modify routes. In future systems, we hope to add other means of travel (bus, airplane), cargos and crews to obtain a system which can be used to solve realistic routing problems.

During the TRAINS-95 evaluation, general criteria for the evaluation of task-based systems were developed. Two parameters were used: time to task completion, and quality of the solution. The quality of the solution was measured in terms of whether the stated goals for a task (routing trains from an initial configuration to a final one) were met, and if they were, how much time was required to complete the planned routes. These criteria have the advantages that they can be applied to any system in which there are objective solution quality measures, and that in many cases the evaluation can be automated. We used these same criteria to evaluate the success of the current TRAINS system.

1.1 Evaluation Goals

A primary goal of TRAINS-95 was to develop a dialogue system sufficiently robust to function despite word recognition errors. The goal of the TRAINS-96 system was to extend TRAINS-95, adding distances and times, and allowing users to modify routes more easily. The evaluation tested these features of the system. The other goals of this evaluation were to:

- identify system deficiencies
- examine the effectiveness of increased robustness in and following the parser (see section 2).

- test the benefit of providing feedback to the user following speech input.

1.1.1 Robustness

In the summer of 1996 an extensive evaluation was made of the sorts of language constructs appearing in TRAINS system utterances that cause fragmentation of utterances into separate speech acts. Approximately 100 dialogues from the TRAINS-95 system were examined for patterns of incorrect fragmentation. Examples include repeated or absent prepositions (e.g. "Let's take the train in from Baltimore to Burlington"), and unnecessary articles (e.g. "Move the train at the Cincinnati to Charleston").

We found two ways of dealing with this unnecessary fragmentation,which leads to incorrect understanding and faulty planning and generation. The first was to add robust rules to the parser. In the TRAINS-95 parser there were some robust rules; we added some more. The TRAINS parser now contains 10 robust rules.

To handle more domain-specific and less well-defined examples, we designed and implemented a template-based post-parser module, patterned after the approaches described in [6, 8]. We describe this module further in section 2 of this paper.

Our intention is that this module should reduce the effects of speech, processing and parsing errors, and that it should enable us to make domain-specific modifications to the parser output before passing it on to the dialog manager. It is possible that some speech acts may be wrongly combined or that some simplifications will lead to incorrect interpretation of utterances. Because the system is currently very simple, we have not seen many of these adverse effects. However, in the future the usefulness of this module may be reduced by improved processing or by increased complexity in the types of conversation the system handles.

1.1.2 Speech Feedback

In the TRAINS system, it is possible to display the output from the speech recognition and/or speech post-processor modules as the user speaks. This shows the subject how his or her speech is being "heard" by the system.

The subject can benefit from speech feedback because it is fairly easy to learn which words the system cannot recognize. Also, this shows the user when he or she is failing to hold the mouse button down while speaking, or is not enunciating clearly. On the other hand, it was our observation that users speak unnecessarily slowly and use overly-simplified language when speech feedback is provided.

1.2 Hypotheses

Our initial hypotheses were:

- The user will interact with the system more naturally and complete tasks faster when feedback is not provided following speech input.

- The post-parser significantly decreases the amount of time spent solving tasks.

2 Post-parser Module

Our post-parser module uses an approach similar to those found in [6, 8]. It comprises three phases, and is designed to allow us to implement more radical or more domain-specific robust parsing techniques than are possible in a general parser.

The post-parser takes as input the logical-form frame structure output by the parser, and outputs the same type of structure, allowing us to maintain system modularity.

The three phases of the post-parser are:

1. Simplification.

2. Speech-act combination.

3. Speech-act identification.

Most modifications that occur are keyed off of the verb in the speech act being processed, because the verb indicates the types of the other objects which may appear in a sentence.

When a user of the TRAINS system makes an utterance, the sound signals pass through Sphinx-II [4] and a sequence of proposed words is output. This sequence is input to the speech post-processor designed by Ringger [5]; the output is a modified sequence of proposed words. This sequence is input to the parser, and a logical form for the utterance is output, along with other

```
( :WANT-NEED
   SA-semantics
   ( :LSUBJ (:DESCRIPTION . :MOVABLE-OBJ) 1.0
     :LOBJ (:DESCRIPTION . :PHYS-OBJ) 1.0
     :LCOMP (:PROP . :PROP) 1.0)
   ()
   ()
   ((lambda (x) (equal (SA-type x) 'SA-TELL)))
   ((lambda (x) (setf (SA-type x) 'SA-ID-GOAL)))
)
```

Figure 1: Example structure for the verb-type *want-need*, showing the template for speech-act combination and the substitution types for speech-act identification

information such as any "noise" words (words that could not be included in the parse) and the probability that the parse is a correct one. Sometimes the output structure is a single speech act; sometimes it is a "compound speech-act" comprising several speech acts. For instance, the utterance "Okay now I want to go to Chicago" parses as a *confirm* speech act and an *id-goal* speech act.

In the first phase of the post-parser, verbs and other parts of the logical form are modified to simplify later processing, not only in the post-parser but also in the dialogue manager. For instance, the verb type *load-with* (as in, "We loaded the boxcar with oranges") is changed to the verb type *load-into* (as in, "We loaded the oranges into the boxcar"). This means that the dialogue manager only has to deal with one *load* verb. The sentence "The train should arrive at Avon" has verb type *arrive.* This is changed to be of type *go-by-path* because in this domain and at this time the sentences "Go from here to Avon" and "the train should arrive at Avon" have the same meaning.

The second phase is template-based. Every verb in the TRAINS domain has a template, telling what classes of objects that verb can take as a subject, direct object, indirect object and/or complement. For example, see figure 1. The *want-need* verb can take any movable object as subject and any physical object as object. If the input to this phase is a compound communications act, and one of the speech acts in that compound communications act is missing one of its parts (as determined by the verb template for that speech act), the other speech acts in the compound communications act are examined to see if any of them has a class corresponding to the missing part. If a suitable speech act is found, the two speech acts are combined.

Figures 2 and 3 show the output from the parser for the sentence "Send it Pittsburgh train to Toronto", and the best output from the post-parser (for the sake of brevity, some parts of the logical form have been replaced by [...]). The sentence "Send it" is incomplete; the verb template for the *move* class allows it to take a complement in the form of a path. The third speech act, "to Toronto", is a path. So the first and third speech acts are combined.

Each speech act output from the parser is classified according to type. This classification determines how the speech act is handled in the dialogue manager. The third phase of the post-parser examines each speech act to determine if its classification matches its verb. If it does not, then the correct classification is substituted. In figure 1, if the verb is of type *want-need* and the speech-act type is *tell*, then the speech-act type is changed to be *id-goal.* For example, "I want to go to Chicago" is parsed as having speech-act type *tell.* The post-parser changes this to *id-goal.*

The post-parser currently has about 80 verb templates, corresponding to the verb types in [1]. Only a very small fraction of these have actually been used and tested, because very few verbs are needed in the current version of the TRAINS system.

Originally, we anticipated adding a fourth phase which would perform some reference resolution, but we believe this is properly the task of the dialogue manager. The post-parser's three phases span the gap between the types of processing performed by the parser and the dialogue manager, enabling each of those modules to maintain generality and improving the understanding of the system in the face of earlier processing errors.

```
(COMPOUND-COMMUNICATIONS-ACT :ACTS
  ((SA-REQUEST :FOCUS NIL :OBJECTS [...]
    :PATHS NIL :DEFS NIL :SEMANTICS
    (:PROP (:VAR :V11935) (:CLASS :MOVE)
(:CONSTRAINT (:AND (:LSUBJ :V11935 :*YOU*) (:LOBJ :V11935 :V11940))))
    [...]:INPUT (SEND IT))
  (SPEECH-ACT :FOCUS NIL :OBJECTS [...]
    :PATHS NIL :DEFS NIL :SEMANTICS :V11961 :NOISE NIL
    :SOCIAL-CONTEXT NIL :INPUT (PITTSBURGH TRAIN))
  (SPEECH-ACT :FOCUS :V11975 :OBJECTS [...]
    :SEMANTICS :V11975 :NOISE NIL :SOCIAL-CONTEXT NIL
    :RELIABILITY 53 [...] :INPUT (TO TORONTO)))
:RELIABILITY 53.25 :MODE TEXT :NOISE NIL))
```

Figure 2: Output from the parser for the sentence "Send it Pittsburgh train to Toronto." ("Send the Pittsburgh train to Toronto" is what the user said; there was a speech recognition error).

```
(COMPOUND-COMMUNICATIONS-ACT :ACTS
  ((SA-REQUEST :FOCUS NIL :OBJECTS [...]
    :SEMANTICS (:PROP (:VAR :V11935) (:CLASS :MOVE)
      (:CONSTRAINT
      (:AND (:LSUBJ :V11935 :*YOU*) (:LOBJ :V11935 :V11940)
      (:LCOMP :V11935 :V11975)))) [...]
    :INPUT (SEND IT TO TORONTO))
  (SPEECH-ACT :FOCUS NIL :OBJECTS [...]
    :PATHS NIL :DEFS NIL :SEMANTICS :V11961 :NOISE NIL
    :SOCIAL-CONTEXT NIL [...] :INPUT (PITTSBURGH TRAIN)))
:RELIABILITY 53.0 :MODE TEXT :NOISE NIL)
```

Figure 3: Output from the post-parser for the sentence "Send it Pittsburgh train to Toronto". The sentence is still somewhat broken up, but what remains can be taken care of by the reference module.

3 Experimental Design

3.1 Experimental Environment

3.1.1 Overview

The experiment was performed over the course of a week and a half in November 1996. Each of the sixteen subjects participated in a session with the TRAINS system which lasted approximately one hour (on average).

3.1.2 Hardware and Software Configuration

All sixteen sessions were conducted in the URCS Speech Lab using identical hardware configurations. The software components used in the experiment included:

- A Sphinx-II speech recognizer developed at CMU [4].
- TRAINS-95 version 2.1 including the speech recognition post-processor [5].[2]
- TrueTalk, a commercial off-the-shelf speech generator (available from Entropics, Inc.).

Subjects, working at a Sun UltraSPARC station, wore a headset with a microphone to communicate with the speech recognizer. While speaking, they held down a button on the mouse. They could also type in a text input window, and click on the map using the mouse.

The TRAINS-96 system communicated with the subjects using the speech generator, by highlighting objects on the map, through a text output window above the map, and by means of dialog boxes.

Figure 4 shows a TRAINS-96 map with a task in progress. Some routes are displayed. A train icon appears at the city of origin, and the city of destination appears highlighted in white (this has since been changed; the destination city now appears with the outline of a train icon). Other highlighted cities, in this case Cincinnati, show places where the train will be delayed.

[2] We ran the TRAINS system on two UltraSPARC stations (speech on one, everything else on the other). The TRAINS system can be run on many architectures, and runs acceptably on a single Sparc10. However, we wanted to obtain natural dialogues having as little time lag as possible between user utterance and system response.

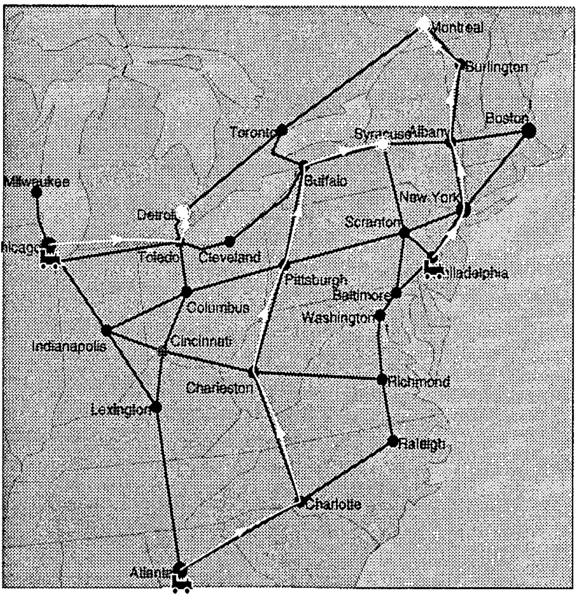

Figure 4: TRAINS-96 map showing a task in progress

Half of the subjects received speech feedback; the other half did not receive speech feedback except while speaking to the system using the practice sentences. About half of the tasks for each subject were performed with the parser robustness in the system turned on; the other half were performed with the robustness turned off (The speech post-processor was on all the time).

3.1.3 Subjects

Of the sixteen subjects, three were recent college graduates, two were high-school students and eleven were undergraduates. All had previous experience using computers and graphical interfaces. None had ever used the TRAINS system before; only four reported ever used any speech recognition system. Four were female, and twelve were male.

3.2 Task Selection

There were five tasks used in the TRAINS-95 evaluation. The routing scenarios for these tasks were designed with the following restrictions:

- Each task involves moving three trains to three cities, with no restriction on which train goes to which city.

- In each scenario, three cities are experiencing delays.

- One of the three routes in each scenario requires more than four hops.

The same tasks were used for the TRAINS-96 evaluation. In addition, we used a sixth task for data collection. In this scenario, the user was given 7 trains at different cities, and had to move as many as possible to the single goal city. There were two restrictions:

- No route was to take longer than 25 hours to complete.

- No segment of track could be used in more than one route.

There were 5 tracks incident on the goal city, and it was possible to move 5 of the trains to the goal city in less than 25 hours.

The data from this last task is not included in the experimental results; it was used only for data collection.

We rotated the first five tasks for each subject, i.e. the first task was used for the first dialogue for the first subject, the second task was used for the first dialogue for the second subject, and so on. The ordering of the tasks for each subject is shown in the tables in Appendix G. The six task was always given to the subject last.

3.3 Procedure

Each subject viewed a 2.5-minute tutorial on a Power PC. The tutorial, which was developed for the experiment, describes how to interact with the TRAINS-96 system using speech and keyboard input, and demonstrates typical interactions using each of these. The subject, therefore, was given some idea of how to speak to the system, but was given no detailed instructions about what could be said. The tutorial simply instructs the subject to "speak naturally, as if to another person." The tutorial also emphasized that the system, not the user, was being evaluated.

3.3.1 Practice

The subject was allowed to practice speech and keyboard input before being given any tasks. At the start of the practice session, the subject was given a list of practice sentences (Appendix A). During this time, all the subjects received speech feedback.

Following the TRAINS-95 evaluation it was suggested that the use of these practice sentences "primes" subjects unnecessarily. Our purpose in using these practice sentences is to make subjects comfortable with speaking to a computer, to allow them to make any slight adjustments in the speed or emphasis of their speech that may be necessary, and to allow us to adjust the input levels of the system so that the subject will have the best possible chance of being understood.

Neither Sphinx-II nor the speech post-processor will understand general speech. Therefore, it would be impossible for us to create a list consisting only of truly domain-independent sentences, although we were able to add some more general statements to this year's practice sentences.

We could analyze our data to find out if subjects did in fact use the form of the practice sentences, but that is not the purpose of these experiments. However, our feeling is that subjects are not unduly "primed" by using these sentences. For example, the practice sentences this year include three questions, but fewer than half of the subjects used questions to help them solve tasks. (This includes task 6, for which the subject had to ask questions to obtain a correct solution.)

Results from the TRAINS-95 evaluation showed that there was a slight learning curve when interacting with the system (see table 1).

Task	*Time*
1	371
2	298
3	203
4	174
5	274

Table 1: Average time to completion per task in the TRAINS-95 evaluation

Therefore, we treated the first dialogue for each subject in this evaluation as a training dialogue. The data from these dialogues are not included in the experimental results.

3.3.2 Task Execution

At the start of each task, the subject was handed a 4"x6" index card with the task instructions. The index cards specified the destinations of the trains and some additional information about cities to be avoided. The exact instructions for each task are provided as Appendix B. The subject did not know the initial locations of the trains until the map was displayed on the computer screen.

Verbal instructions given to the subject were:

- Take your time reading the task card, but once you have started the plan, try to work quickly.
- You may speak or use the mouse or keyboard, but please try not to use the mouse or keyboard unless you feel the system is really not understanding you.
- The experimenter cannot answer questions.

3.3.3 Questionnaires

After each task, the subject was given a questionnaire to complete. This asked if the subject had difficulty completing the task, and if so, what the the subject thought the causes of that difficulty were. We did this to test if the subject could differentiate between system performance without the robust parsing capabilities and system performance with them. Some subjects also noted any specific difficulties they encountered. The questionnaire is provided in Appendix D; the responses are provided in Appendix F.

After completing the final task, the subject completed a more general questionnaire. This designed to give us some background information and to allow the subject to comment about system performance in general. This questionnaire is provided in Appendix C, and the responses in Appendix E.

4 Experiment Results

Metrics were collected for each subject. Appendix G contains tables detailing the raw data collected. The figures and statistics in this section summarize the data.

Of the sixty-four dialogues included in the results, there were seventeen in which the stated goals were not met (this figure does not include dialogues where the system crashed). In four of these, the subject thought he or she had met the goals. In seven, the subject did meet all the goals at some point in the dialogue, but in the final configuration the goals were not met. The subject tried to alter one or more routes, and in the process failed to meet one or more goals. In the other six, the subject did not meet all the goals at any point in the dialogue.

The system crashed in five of the dialogues included in the results. In addition to these dialogues, the system crashed in three dialogues, but did so very early in the dialogue, and so we allowed the subject to start over. The resulting dialogues are marked with a *.

Eight of the subjects used the keyboard, four of them a significant amount (more than five times in at least one dialogue). Four subjects used the mouse, none a significant amount.

4.1 Task performance results

Results are also given for the average completion length in miles, but the results for this metric are not as significant as the results for time to completion. Generally, if a subject completed a task, the solution resembled the solutions of other subjects for that task to a large degree.

Figures 5 and 6 show the average time for a dialog per task, in seconds, and the average length of the solution, in miles. The pale bar gives results when the dialogs were conducted with robustness turned off. The darker bar shows the results when the dialogs were conducted with robustness turned on.

The time to completion in four of the five tasks is lower with robustness. In the fifth task it is higher. This result is due to the large amount of time subject eight spent on task five. Because our sample size is small, any very unusual data can skew the results.

When robustness was turned on, the length of the solution is longer in four of the five tasks. However, the differences are extremely slight. Also, most subjects attempted simply to complete the task, not to obtain an optimal solution.

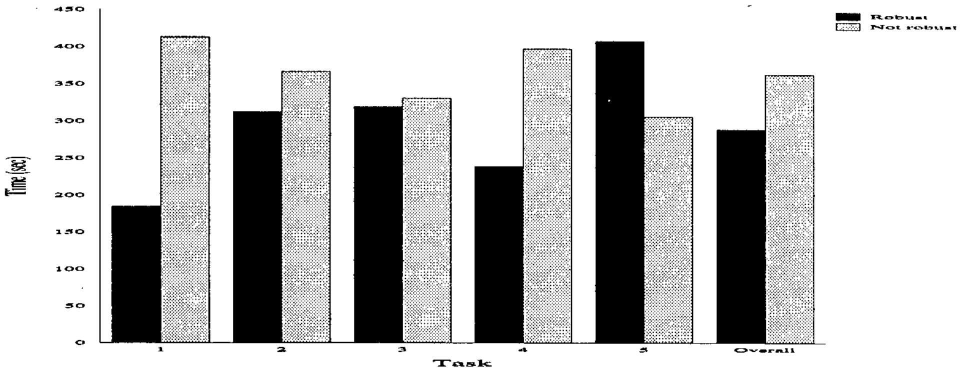

Figure 5: Average time to solution; robustness factor only

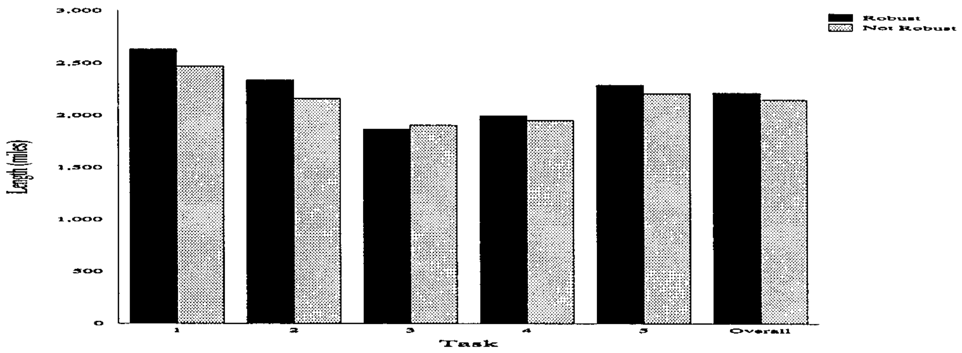

Figure 6: Average length of solution; robustness factor only

Figure 7 shows the average time for a dialog per task; results for dialogs where speech feedback was supplied are the darker bars, and results for dialogs without speech feedback are the lighter bars. Figure 8 shows the average length of solution per task, with and without speech feedback.

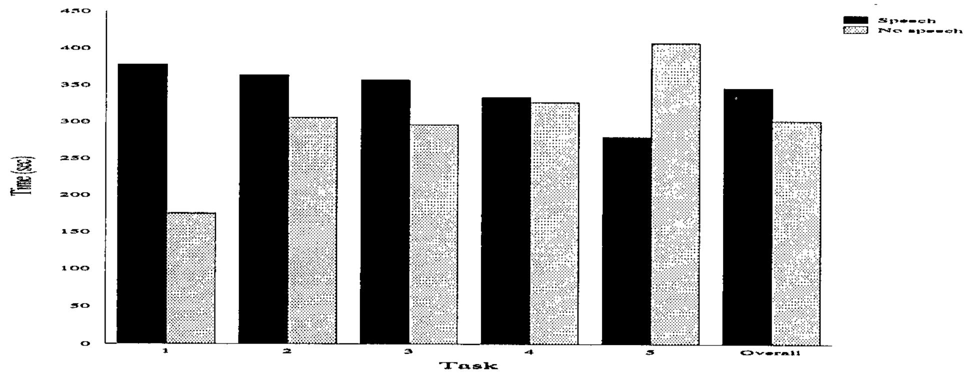

Figure 7: Average time to solution; speech factor only

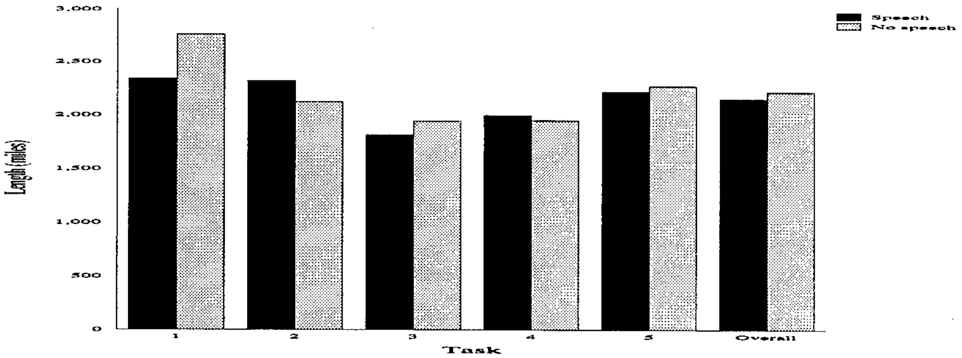

Figure 8: Average length of solution; speech factor only

The time to completion in four of the five tasks is lower without speech feedback. The results for the fifth task are different. Again, this is due to the large amount of time subject eight spent on task five.

In three of the tasks, the length of the solution was less when speech feedback was provided. Again, these results are not statistically significant.

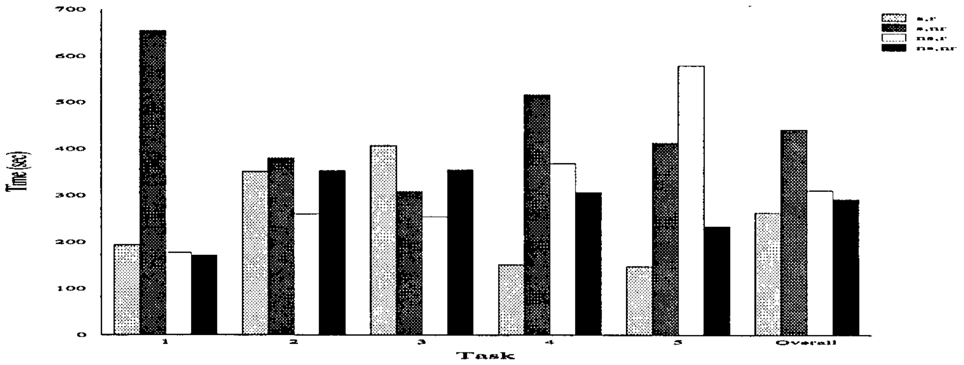

Figure 9: Average time to solution

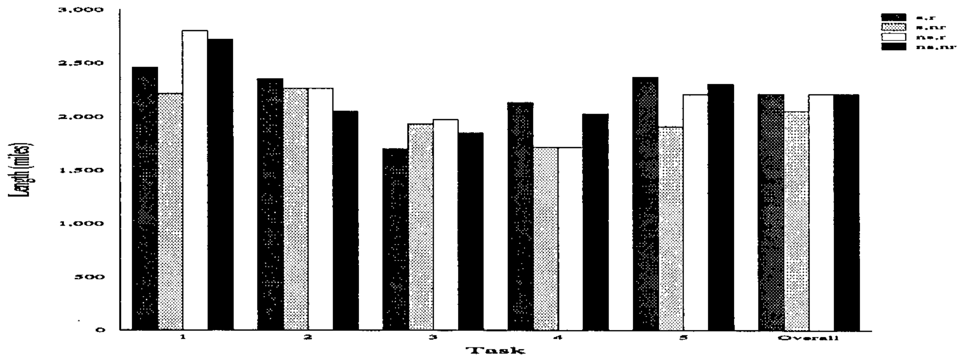

Figure 10: Average length of solution

Figure 9 shows the average time for a dialog, per task. Results are given for tasks in which:

- both robustness and speech feedback were used (s,r)
- speech feedback was provided, but robustness was not used (s, nr)
- robustness was used, but speech feedback was not provided (ns, r)
- neither robustness nor speech feedback were used (ns, nr)

Figure 10 divides the data in the same way as figure 9, but shows results for the length of the routes.

In two of the tasks the time to completion is lowest when the robust parts of the parser are being used and there is no speech feedback. In another two the time to completion is lowest with the robustness and the speech feedback. Overall, the best times were obtained when both speech feedback and robustness were used.

4.2 Subject response results

The following tables summarize the subjects' responses to questionnaire B. The subjects were asked to estimate the contribution of three parts of the system to the difficulties they experienced in completing the tasks:

- One: speech recognition
- Two: language understanding.
- Three: route planning.

Table 2 gives the average response to each question. Table 3 compares responses where the subject had speech feedback to responses where speech feedback was not provided. Table 4 compares responses where robustness was used to responses where robustness was not used. In all cases, the data is divided on a per-task basis.

Task	*One*	*Two*	*Three*
1	4.12	4.73	3.63
2	5.27	5.23	3.92
3	5.31	5.77	2.77
4	5.46	5.31	3.42
5	4.86	5.83	3.75
Overall	5.04	5.38	3.49

Table 2: Average contribution to difficulty per task

On the whole, subjects were less likely to blame the route planner than they were to blame the language understanding parts of the system.

	With Speech			*Without Speech*		
Task	*One*	*Two*	*Three*	*One*	*Two*	*Three*
1	5.08	5.58	4.67	2.96	3.87	2.58
2	6.79	6.43	3.57	3.50	3.83	4.33
3	6.83	6.33	2.00	4.00	5.29	3.43
4	6.07	6.21	3.57	4.75	4.25	3.25
5	5.40	5.80	4.00	4.42	5.86	3.57
Overall	6.08	6.10	3.55	3.96	4.68	3.44

Table 3: Average contribution to difficulty per task: speech factor only

When the speech feedback was on, subjects were more likely to blame the natural language parts of the system for difficulties in completing tasks than they were when it was off.

	With Robustness			*Without Robustness*		
Task	*One*	*Two*	*Three*	*One*	*Two*	*Three*
1	3.80	4.03	2.71	4.50	5.70	4.90
2	6.50	5.93	3.79	3.83	4.42	4.08
3	5.43	5.93	2.21	5.17	5.58	3.42
4	4.50	4.30	3.60	6.06	5.94	3.31
5	4.58	5.64	2.79	5.20	6.10	5.10
Overall	5.04	5.22	2.98	5.03	5.55	4.05

Table 4: Average contribution to difficulty per task: robustness factor only

When the robustness was off, subjects were more likely to blame the route planner for difficulties they may have encountered than they were when it was on. They were not otherwise able to differentiate between system performance with and without the parser robustness (they were not told that we were varying system behavior during the evaluation).

5 Discussion

Our preliminary results indicate that tasks are completed more quickly when the robustness in the parser is on. There is a smaller difference in perfor-

	Task 1		*Task 2*		*Task 3*	
	Mean	*St. Dev.*	*Mean*	*St. Dev.*	*Mean*	*St. Dev.*
Overall	267.36	235.16	336.85	156.62	324.31	231.83
Speech	377.80	292.65	363.29	167.07	356.83	236.07
No speech	175.33	106.67	306.00	137.18	296.43	224.41
Robust	184.43	107.24	311.71	169.06	318.86	218.13
Not robust	412.50	314.40	366.17	134.94	330.67	246.69

	Task 4		*Task 5*	
	Mean	*St. Dev.*	*Mean*	*St. Dev.*
Overall	330.42	229.87	356.3	383.68
Speech	333.83	222.00	280.00	141.27
No speech	327.00	237.43	407.17	474.95
Robust	237.80	228.21	407.20	522.90
Not robust	396.57	207.08	305.40	125.75

Table 5: Means and standard deviations per task for single factors

mance between tasks completed with speech feedback and those completed without it. Unfortunately, while the means indicate perceptible differences, the standard deviations are also large (see table 5).

There are two causes for the large standard deviations. The first is that our sample size is very small. An experiment like this should be performed with a minimum of 100 subjects. Our goal in this evaluation, however, was obtain indications of the correctness of our hypotheses rather than to demonstrate their correctness beyond all possible doubt. The second cause is the amount of time spent altering routes in some dialogues. Some of the difficulty subjects experienced in altering routes was caused by a bug in the problem solver, which has since been fixed.

We conducted an anova test of the data. The F-critical values were:

- for robustness: 1.353
- for speech: 0.236
- for robustness and speech: 1.471

These indicate that our results are not statistically significant for any of the variables in the experiment.

The one very clear result of this evaluation is that it is still difficult to modify routes in the TRAINS system. As has already been noted, in seven dialogues the subject completed the task, and then as a result of trying to modify one or more routes partially or completely undid that solution. Even in cases where the task was completed, much time was often spent modifying routes. For instance, subject three spent almost nine minutes on task two. After two minutes, the task was completed. The other seven minutes were spent attempting to modify one of the routes.

In some cases the difficulties arose because the subject did not speak naturally to the system. (Some subjects said things like, "Send train Chicago to Toledo.") In other cases problems were caused because the subject tried to modify a route that was not the current focus of the discourse, and did so without using language cues such as "now" and "instead." Problems also arose because of speech recognition errors. However, in most cases the fault lies with the discourse manager or planner. The subject would try five or six different ways of asking for a modification, and the system would simply refuse to carry it out.

When over two-thirds of the time spent solving some tasks is spent in modifying routes rather than in completing the task itself, any other factors being considered will be over-shadowed. Nonetheless, we do see indications that our initial hypotheses were correct. In situations where there are speech recognition errors and previous processing errors, a robust post-parsing module can improve the time to completion of tasks. Also, providing users with speech feedback may adversely affect their performance on tasks.

From the answers subjects gave to questionnaire B, we can see indications that they could not tell any differences in performance of the system when the robustness is turned on. This may be because the frustration of trying and failing to modify routes blinds the user to other aspects of system performance.

Subjects were more likely to blame the parser and natural language understanding parts of the system when they received speech feedback than when they did not. Perhaps the subjects had difficulty separating the different modules of the system unless one module's performance was made obvious (as in the case of speech feedback). This is a positive result, because we want users of the system to think of it as a single intelligent agent.

6 Future work

An important possible future direction for work with the data from this evaluation is a closer examination of the dialogues to determine the amount of time spent modifying routes, and to explore ways in which we can make this easier for users of the TRAINS system. In addition, the data from this evaluation is being used to evaluate the TRAINS speech post-processor.

We have concluded that in future evaluations we need a larger number of subjects in order to get meaningful results. We have also decided that rotating the order in which tasks are performed introduces unnecessary complications into the experimental design. If the tasks are rotated in a future evaluation, one task should be reserved for practice and only the others should be rotated. This will facilitate evaluation of the results.

7 Acknowledgments

We would like to thank James Allen for his expert assistance in planning and carrying out this evaluation, George Ferguson for modifying the system to allow us to perform our experiments and for correcting many system errors, and Mike Tanenhaus and Joy Hanna for their advice in analyzing our results.

A TRAINS Practice Sentences

The following are sentences you can use to practice speaking and typing to the TRAINS system. Feel free to alter the position of your headphones and microphone or the volume levels of the system. You may ask for help at this stage. Take as much time as you want to practice.

- Send the train from Atlanta to Philadelphia.
- Go to Chicago via Pittsburgh.
- How are you?
- No, go through Atlanta instead.
- Yeah, so are we done?
- Good.
- There is heavy traffic there.
- The train at Toronto should go to New York via Buffalo and Syracuse.
- Don't go through Scranton.
- And then up to Richmond.
- Where are the trains?
- Cancel that.
- I want to work fast.
- I'm done.
- How far is that route?

B Task Instructions

B.1 Task 1

Perform the following task as quickly as possible:

Montreal, **Boston** and **Chicago** each need a train to be moved there. Lake-effect snow is causing delays in Buffalo. Heavy traffic in Central and Southern Ohio is causing delays and should be avoided.

When you have completed your plan, or if you want to give up, inform the system that you are done.

B.2 Task 2

Perform the following task as quickly as possible:

You need to construct a plan to get one train to **Toronto**, one to **Lexington** and a third to **Atlanta**. Expect delays through Indianapolis and Columbus due to bad weather. Heavy traffic in Detroit is also causing delays there.

When you have completed your plan, or if you want to give up, inform the system that you are done.

B.3 Task 3

Perform the following task as quickly as possible:

You need to plan optimal routes to move trains to **Albany**, **Raleigh** and **Lexington**. Much of New York State has been paralyzed by freezing rain, so routes through New York should be avoided as much as possible.

When you have completed your plan, or if you want to give up, inform the system that you are done.

B.4 Task 4

Perform the following task as quickly as possible:

Trains are needed at **Lexington**, **Philadelphia** and **Washington** in as little time as possible. Traffic congestion in New York City is causing delays there. Additonally, the areas close to Pittsburgh and Buffalo are experiencing heavy snowfall and long delays are likely through those cities.

When you have completed your plan, or if you want to give up, inform the system that you are done.

B.5 Task 5

Perform the following task as quickly as possible:

You need to find optimum routes to get one train to **Milwaukee**, one to **Lexington** and a third to **Washington**. Heavy storms are causing delays in Baltimore and Eastern Pennsylvania.

When you have completed your plan, or if you want to give up, inform the system that you are done.

B.6 Task 6

Perform the following task as quickly as possible:

Your goal is to move as many trains as you can to **Scranton**. There is heavy traffic in Cincinnati, and an insurrection is taking place in Toronto. New York is experiencing bad weather.

No segment of track may be used twice.

You must plan so that no route takes longer than 25 hours.

When you have completed your plan, or if you want to give up, inform the system that you are done.

C TRAINS Questionnaire A

Thank you for participating in the 1996 TRAINS system evaluation. To help us better understand the experiment, please answer the following questions:

Name:

1. Have you ever used the TRAINS system before?
2. Have you ever used a speech recognition system before?
3. If you used the keyboard mode, did you use it because:
   (a) The system had trouble understanding you?
   (b) It allowed you to solve the problem more quickly?
   (c) It was more fun to use?
   (d) Other:
4. What would you suggest to make the system more effective? You may use the back if you don't have enough space.

D TRAINS Questionnaire B

Name:

Dialogue number:

Please answer the following question, considering only the dialogue which you have just completed.

If you had difficulty, what reason did you think was mostly the cause? Rate each as very important to not at all.

1. The system had too many speech recognition errors.

   very important — not important

2. The system had trouble understanding the language in general.

   very important — not important

3. The system is not a very good route planner.

   very important — not important

E Questionnaire A Responses

Subject experiences

Subject	*Previous TRAINS Use*	*Previous SR Use*	*Keyboard Used?*	*Easier To Use*	*Faster*	*More Fun*	*Other*
1	No	No	No				
2	No	No	Yes	Yes	Not Necessarily	No	
3	No	No	Yes	Yes			
4	No	Yes[1]	No				
5	No	No	No*				
6	No	No	No				
7	No	Yes[2]	Yes				
8	No	No	Yes	Yes			Yes[3]
9	No	No	Yes	Definitely	Yes	No	
10	No	Yes	No				
11	No	No[4]	Yes	Yes			
12	No	No	Yes	Yes	Not Really	No	
13	No	No	Yes	Yes			
14	No	No	No				
15	No	No	No				
16	No	No	No				

∗ - Actually, the subject did use the keyboard for one utterance.

[1] - very primitive

[2] - PlainTalk on a Power Macintosh

[3] - I couldn't pronounce the city

[4] - Not really - well actually a Power Mac that tells knock-knock jokes

Suggestions for the TRAINS system

Subject	*Suggestions*
1	It seems unclear on which train it is supposed to use and confuses them.
2	I had trouble modifying my proposed routes ... I also had trouble saying "I would like to take the Boston train to Pittsburgh via New York." It was usually witty, if you don't mind being called bozo or your idiocy.
3	see dialogue 6.
4	I just had some trouble switching between trains and getting the trains to move backward.
5	There needs to be a way to select routes that have been set in the past. When I finished doing one route then did another and wanted to cancel the first I could not do it.
6	I'm not sure what could be done to fix it but if my instructions started to slur the machine didn't recognize what I was saying. An easy way to verbally cancel a move you did awhile ago would also be nice.
7	I had some problems convincing TRAINS to let me change routes, either because I realized that a better path was available or because the system picked the worst possible route to use. I also had some difficulty with getting TRAINS to understand me - it kept thinking I had said other things, so the language recognition was a little off for me.
	Touch up on selecting which train, because often I would specify a train, but [it] would only use the previous one, even using keyboard.
	Maybe have the computer prompt the speaker when a route doesn't work i.e. Sp.: Move the train from A to B. Computer: confusion insults etc. Would you like me to try another route? Sp.: Yes please
11	It had trouble understanding certain commands for no apparent reason. It seemed difficult to undo a command once it was done.
14	I don't know.
16	The error messages, although clear, were not very helpful. The system should permit a user to edit a sentence they have said. The system should ask a few more questions, and it should ask if a route should be locked so it won't modify it again without an explicit unlock statement.

Some users made comments not only on questionnaire A, but also on questionnaire B. These suggestions appear with the answers to questionnaire B.

*Note that some users have a very machine-oriented view of the system. This showed itself in other ways too. There were two subjects who tried to speak without using articles, e.g. "Move train to Chicago." or even "Montreal Chicago."

F Questionnaire B Responses

Where there are blanks, the subject did not give any value. The values range from 1 to 10.

Perceived cause of difficulties				
Subject	*Dialog*	*Speech Recognition Errors*	*Poor Language Understanding*	*Poor Route Planning*
1	2	8	8	1
1	3	8	8	1
1	4	9	9	1
1	5	8	8	1
2	2	7	7	1
2	3	9	7	7
2	4		10	3
2	5		6	2
3	2	5	7	5
3	3			
3	4	4	4	8
3	5	5	5	7
4	2	1.5	6.5	3.5
4	3	1.5	3.5	3.5
4	4	1.5	5.5	4.5
4	5	2	3.5	3.5
5	2	8	7	3
5	3	3	3	3
5	4	9	7	3
5	5	7	8	4
6	2	3	3	5
6	3	6	5	4
6	4	2	1	1
6	5	5	4	4

Perceived cause of difficulties				
Subject	*Dialog*	*Speech Recognition Errors*	*Poor Language Understanding*	*Poor Route Planning*
7	2	3	2	1
7	3	4	4	7
7	4	9	7	7
7	5	2	2	4
8	2	7.5	5.5	3.5
8	3	8.5	9.5	5.5
8	4	6.5	6.5	3.5
8	5	7.5	4.5	4.5
9	2	1	4	2
9	3	5	8	10
9	4	8	7	4
9	5	8	8	4
10	2	1	1	2
10	3	1	1	3
10	4	1	1	1
10	5	2	2	2
11	2	8	8	2
11	3	7	7	2
11	4	7	7	2
11	5	3	3	3
12	2	5	7.5	5.5
12	3	6	8	4
12	4	4.5	6	4
12	5	3.8	4.2	2.5

Perceived cause of difficulties				
Subject	*Dialog*	*Speech Recognition Errors*	*Poor Language Understanding*	*Poor Route Planning*
13	2	6.5	4.5	5
13	3	6	7	7
13	4	6.5	4.5	2
13	5	7.5	7	7
14	2	5	4	3
14	3	2	2	2
14	4	6	7	4
14	5	3	5	5
15	2	5	8	1
15	3	8	7	1
15	4	6	6	1
15	5	4	4	1
16	2	2	2	5
16	3	4	8	4
16	4	2	2	2
16	5	2	1	2

Suggestions for the TRAINS system	
Subject	*Suggestions*
3	[6] The train was not allowed to change plans and if you tried one route and then switched it, it would try and use that path and the new one, so it went out of its way a lot.
4	[1] Couldn't get the trains to go back to the cities they came from to try a different path.
	[2] Still couldn't get the trains to move back to original cities.
	[4] Couldn't get the train from Charlotte back.
	[6] Trouble switching between trains.
7	[2] I had very little difficulty with this trial.
	[3] The main problem I had was convincing TRAINS to change routes.
	[4] Inability to cancel badly-planned routes and difficulty in getting TRAINS to understand me.
	[5] Very little difficulty.
8	[3] I could not move a train I had left for awhile.
	[5] Still had small problem with unused engines.
9	[3] The computer refused to allow me to move the train from New York to Boston. Check transcript.
16	[5] It works better if it gets simple instructions. Unlike a person.

The dialogue number, if any, with which a comment is associated, appears in brackets before the comment itself.

G Data for task six

We used task six primarily for data-collection purposes. Only 2 subjects completed this task and satisfied all the constraints. Several other subjects almost completed the task, or completed it and then went on and added other cities which caused some segment of track to be used more than once. Very few subjects asked any questions about the lengths of routes or the time they were taking. In fact, those subjects who did complete the task or almost complete the task did so by pure luck, since they didn't ask questions and therefore could not know that they had satisfied all the constraints.

For this task, the robustness was on for all subjects. Speech feedback was given only to those subjects who had speech feedback in earlier dialogues.

A summary of the data for each subject on this task is given below. The questions referred to are those in questionnaire B.

Task six data

Subject	*Time*	*Completed (y/n)*	*Question 1*	*Question 2*	*Question 3*
1	399	n	8	8	1
2	450	n	8	8	4
3	501	n	3	3	6
4	1870	n	2.2	5.2	3.5
5	530	y	7	5	3
6	691	n	5	5	3
7	896	y	7	5	8
8	696	n	4.5	8.5	4.5
9	606	n	9	8	2
10	211	n	1	1	2
11	448	n	8	4	2
12	740	n	7	7	4.2
13	162	n	4	4.5	5
14	753	n	3	4	3
15	894	n	7	7	1
16	344	n	3	6	1
Average	646.94		5.42	5.58	3.33

The average time spent on a dialogue about this task is almost twice as long as the average time spent on any earlier task. The subjects did not think that the natural language processing or problem solving capabilities of the system contributed to the difficulty of this task any more than to the difficulty of earlier tasks.

This task is not very difficult, although speech recognition errors can make it boring to complete. In some cases, the subject did not think the system could understand questions because he or she had tried to ask questions in earlier dialogues and had not been understood. Most of the subjects simply did not read the scenario carefully enough.

H Data Collected

This appendix contains the data collected for the sixty-four dialogues included in the TRAINS-96 system evaluation. The first table gives a summary of the average time to completion for each subject, with and without robustness. The following tables give the data for each task. The dialog number refers to the relative position of the task for that subject. For example, the first task formed the fifth dialogue for subject two. If a dialogue is marked with an asterisk (*), then the system crashed near the beginning of that dialogue and the subject was allowed to start it again.

	Speech	*With*	*Without*	
Subject	*(yes/no)*	*robustness*	*robustness*	*Overall*
1	yes	191.5	375.0	252.7
2	no	277.5	545.0	366.7
3	yes	524.0	405.0	464.5
4	no	122.5	459.0	290.8
5	yes	548.0	488.0	518.0
6	no	76.5	176.0	126.3
7	yes	139.5	218.5	179.0
8	no	911.5	374.0	642.8
9	yes	339.5	770.5	555.0
10	no	106.5	172.5	139.5
11	yes	203.0	395.5	299.3
12	no	410.0	446.5	428.3
13	yes	164.0	513.5	338.8
14	no	214.5	86.5	150.5
15	yes	127.5	194.0	149.7
16	no	371.0	207.0	289.0

Table 1: Average time to completion for each subject

Task 1						
	Dialog	*Robustness*	*Met*	*Completion*	*Route*	*Route*
Subject	*Number*	(yes/no)	*Goals?*	*Time (sec.)*	*Length (m.)*	*Time (hrs.)*
2	5	yes	yes	326	3274	108
3	4	no	yes	405	1627	53
4	3	yes	no	107		
5	2	yes	yes	373		
7	5	yes	yes	75	2457	68
8	4	no	yes	309	2457	68
9	3	no	no	903		
10	2	yes	yes	133	2329	72
12	5	yes	no	144		
13	4	yes	yes	133	2457	68
14	3	no	yes	33	2977	87
15	2	no	crashed			

Task 2						
	Dialog	*Robustness*	*Met*	*Completion*	*Route*	*Route*
Subject	*Number*	(yes/no)	*Goals?*	*Time (sec.)*	*Length (m.)*	*Time (hrs.)*
1	2*	yes	yes	127	2264	67
3	5	yes	no	524		
4	4	no	no	540		
5	3	no	yes	229	2264	67
6	2	yes	no	93		
8	5	yes	yes	371	2264	67
9	4	yes	yes	551	2264	67
10	3	no	yes	214	1835	59
11	2	no	yes	363	2264	67
13	5	no	no	548		
14	4*	yes	no	315		
15	3	yes	yes	201	2534	82
16	2	no	yes	303	2264	67

Task 3

Subject	*Dialog Number*	*Robustness* (yes/no)	*Met Goals?*	*Completion Time (sec.)*	*Route Length (m.)*	*Route Time (hrs.)*
1	3	yes	yes	256	1931	64
2	2	yes	yes	229	2046	62
4	5	yes	yes	138	1941	57
5	4	yes	yes	723	1465	53
6	3	no	yes	211	1849	58
7	2	no	yes	91		
9	5	no	yes	638	1941	57
10	4	yes	yes	80	1931	64
11	3	yes	no	239		
12	2	no	no	710		
14	5	no	yes	140		
15	4	no	yes	194	1919	61
16	3	yes	no	567		

Task 4

Subject	*Dialog Number*	*Robustness* (yes/no)	*Met Goals?*	*Completion Time (sec.)*	*Route Length (m.)*	*Route Time (hrs.)*
1	4	no	no	375		
2	3	no	no	545		
3	2	no	crashed			
5	5	no	no	747		
6	4	yes	yes	60	1715	52
7	3	yes	yes	204		
8	2	no	yes	439	2654	79
10	5	no	yes	131	1715	52
11	4	no	yes	428	1715	52
12	3	yes	no	676		
13	2	yes	yes	195	2554	86
15	5	yes	yes	54	1715	52
16	4	no	yes	111	1715	52

Task 5						
Subject	*Dialog Number*	*Robustness* (yes/no)	*Met Goals?*	*Completion Time (sec.)*	*Route Length (m.)*	*Route Time (hrs.)*
1	5	yes	crashed			
2	4	yes	crashed			
3	3	yes	crashed			
4	2	no	yes	378	2415	77
6	5	no	yes	141	2804	78
7	4	no	yes	346	1909	61
8	3	yes	no	1452		
9	2	yes	yes	128	2175	71
11	5	yes	yes	167	2568	77
12	4	no	yes	183	1703	53
13	3	no	no	479		
14	2*	yes	yes	114	2595	80
16	5	yes	yes	175	1823	55

References

[1] James F. Allen. Logical form in the trains-96 system. Technical Report to appear, University of Rochester, 1996.

[2] George Ferguson, James Allen, and Brad Miller. TRAINS-95: Towards a mixed-initiative planning assistant. In *Proceedings of the Third Conference on Artificial Intelligence Planning Systems*, 1996.

[3] George M. Ferguson, James F. Allen, Brad W. Miller, and Eric K. Ringger. The design and implementation of the TRAINS-96 system: A prototype mixed-initiative planning assistant. Technical Report 96-5, University of Rochester, October 1996.

[4] D. Huang, F. Alleva, H.W. Hon, M. Y. Hwang, K.F. Lee, and R. Rosenfeld. The Sphinx-II speech recognition system: An overview. *Computer, Speech and Language*, 1993.

[5] Eric K. Ringger and James F. Allen. A fertility channel model for post-correction of continuous speech recognition. In *Proceedings of the 1996 International Conference on Speech and Language Processing*, October 1996.

[6] Stephanie Seneff. A relaxation method for understanding spontaneous speech cutterances. In *Proceedings of the 1992 DARPA Speech and Natural Language Workshop*, February 1992.

[7] Teresa Sikorski and James Allen. A task-based evaluation of the TRAINS-95 dialogue system. In *Proceedings of the ECAI Workshop on Dialogue Processing in Spoken Language Systems*, August 1996.

[8] David Stallard and Robert Bobrow. Fragment processing in the Delphi system. In *Proceedings of the 1992 DARPA Speech and Natural Language Workshop*, February 1992.